

# 第三篇 流行病学中的统计方法

## 第十五章 遗传学中的统计学

李照海<sup>1</sup> 谢民育<sup>2</sup> 许宗利<sup>3</sup>

<sup>1</sup> 乔治华盛顿大学 <sup>2</sup> 华中师范大学

<sup>3</sup> 中山大学公共卫生学院医学统计学教研室

### 第一节 基本概念

分子遗传学的最新发展给人类复杂性状的遗传研究提供了机遇。很多人类疾病,如囊性纤维化病、胰岛素依赖性糖尿病、高血压,以及精神分裂症都被认为有遗传成分,确定可能影响此类疾病的基因位置对病因研究极为重要,且可能会导致更好的治疗方法。本章的目的是对遗传数据的分离分析和连锁分析作一个介绍,我们首先从基本的遗传概念和相关术语开始。

#### 一、遗传术语

每个人有 23 对染色体。其中一对是性染色体(sex chromosomes),它由染色体 X 和 Y 组成,女性为 XX 型染色体,男性为 XY 型染色体,其余的 22 对称为常染色体(autosomal chromosomes),在本章我们主要研究常染色体。一对染色体可想象为两条平行的直线。染色体上一个给定的位置(好比两条平行直线上的一段或一点)叫做基因座(locus),在基因座上不同形式的遗传性变量叫做等位基因(alleles)。在分子遗传学或生物学中,基因(gene)这个术语既指等位基因又指基因座,等位基因经常用字母如 A、a、B、b 和数字 1、2、3 等来表示。在一群体中,一给定基因座上某一等位基因的比率叫做该等位基因的频率或概率(allele frequency, allele probability),例如,  $P = P(A) = 0.3$ , 表示在给定基因座上有 30% 的等位基因是 A, 基于遗传数据,我们可以估计等位基因的频率。在任一给定的基因座上,每个人有两个等位基因,例如 AA、Aa 或者 aa, 基因座上这样一对等位基因叫做基因型(genotype),等位基因的顺序并不影响基因型的类型,即可认为“Aa”和“aA”为同一基因型。如果基因座上的两个等位基因相同,如 AA 和 aa, 则称此基因型为纯合的(homozygous)。如果不同,则称此基因型是杂合的(heterozygous)。通常,一个基因座上基因型的确定是困难的,然而,基因型表现出来的某些特征却是容易观察到的,例如眼睛的颜色和身高,我们称观察到的这些特征为基因型的表现型(phenotype)。基因型和表现型之间的

关系并不一定是一一对应的,可能几种不同的基因型对应同一种表现型,例如 AA 和 Aa 有同样的表现型,但和 aa 的表现型不同,此时我们说 A 相对于 a 是显性的(dominant),或者说 a 相对 A 是隐性的(recessive),AA 的表现型称为显性的,与 aa 相应的表现型称为隐性的;如果基因型 Aa 对应的表现型既不同于 AA 的表现型,也不同于 aa 的,就称 A 和 a 是共显性的。例如,在决定人类血型的基因座上有三个等位基因:A、B 和 O,基因型 AA 和 AO 有同样的表现型,即 A 型血,基因型 BB 和 BO 有同样的表现型 B 型血,因此,A 相对于 O 是显性的,B 相对于 O 也是显性的,基因型 OO 有隐性表现型 O 型血,基因型 AB 有共显性的表现型 AB 型血。

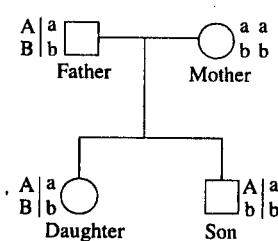


图 15-1 单体型示意图

当我们同时考虑多个基因座时,个体从父亲或母亲那儿获得的等位基因(位于不同基因座上)叫做单体型(haplotype),一对单体型构成了联合基因型。假设有两个基因座,基因座 1 有两个等位基因 A 和 a,基因座 2 有 B 和 b,图 15-1 提供了一个有父亲、母亲,一个儿子和一个女儿的家庭,在此家庭中,儿子从父亲处获得单体型 Ab,从母亲处获得单体型 ab;女儿从父亲处获得单体型 AB,从母亲处获得单体型 ab。图 15-1 中的方框和圆圈分别代表男性和女性。注意到在图 15-1 中某些个体的两

种单体型之间有一垂直长条,长条同一边的等位基因均来自父亲或均来自母亲,即每个单体型是源于父亲还是母亲是清楚的。儿子从父亲处得到的单体型 Ab,它不同于父亲的两个单体型 AB 和 ab,这表明在父亲传递基因给儿子的过程中,两个基因座上的等位基因发生了重组(recombined),这种现象叫做交换(cross-over),只有当两个基因座之间发生奇数次交换时,重组才可观察到。两个基因座上发生奇数次重组的概率叫做重组率(recombination fraction),用  $\theta$  来表示。

孟德尔第一定律(Mendel's first law)或独立分离原理(the principle of independent segregation)是指个体以相等的概率获得父母基因型中的两个等位基因中的一个,假设父亲(或母亲)的基因型为 Aa,则上述原理表明  $P\{\rightarrow A|Aa\} = P\{\rightarrow a|Aa\} = 1/2$ ,这里  $\{\rightarrow A|Aa\}$  表示已知父亲(或母亲)的基因型为 Aa 的条件下,其传递等位基因 A 给后代这一事件。

如果两基因座位于不同的染色体上,则两个基因座上的等位基因的传递是相互独立的,例如,假设父亲(或母亲)有基因型 AaBb(两基因座上的基因型),且两个基因座在不同的染色体上,那么,

$$P\{\rightarrow AB|AaBb\} = P\{\rightarrow A|Aa\} \times P\{\rightarrow B|Bb\}$$

这种自由组合(independent assortment)的原理通常称为孟德尔第二定律(Mendel's second law)。

## 二、Hardy-Weinberg 平衡

随机婚配是指任何一位女性与任何一位男性婚配的等可能性,从遗传学的角度来说,婚配类型的概率等于女性和男性基因型概率的乘积,例如: $P\{AA \times Aa\} = P(AA)P(Aa)$ ,符号 AA × Aa 表示 AA 型个体与 Aa 型个体相婚配。

接下来,我们考虑有两个等位基因 A 和 a 的基因座,假定此两个等位基因频率分别

为,  $P(A) = p$ ,  $P(a) = q$ , 这里  $p + q = 1$ 。如果群体中三种基因型的频率分别为:

$$P(AA) = p^2, P(Aa) = 2pq, P(aa) = q^2 \quad (1)$$

则称此群体处于平衡态(equilibrium)。在随机婚配的条件下, 子代的等位基因频率和基因型频率与父代的相同, 即对子代来说下面式子仍成立:

$$P(AA) = p^2, P(Aa) = 2pq, P(aa) = q^2$$

且

$$P(A) = p, P(a) = q$$

对于一个处于平衡态的群体, 它们世代相传且基因型频率和等位基因频率保持不变, 这就是 Hardy-Weinberg 定律(Hardy(1908); Weinberg(1908))。

如果当前一代的基因型频率不满足条件(1), 则经过随机婚配以后, 下一代就达到了平衡(Wentworth 和 Remick (1916))。关于 Hardy-Weinberg 定律的详细内容, 读者可参看 CC Li 的关于群体遗传学的优秀专著(Li(1988))。

### 三、连锁和连锁平衡

从孟德尔第二定律我们可以得出: 如果两个基因座在不同的染色体上, 则两个基因座上的等位基因的传递(分离)是相互独立的, 即重组率为  $\theta = \frac{1}{2}$ ; 如果同一染色体上两个基因座相临近, 则同源于父亲(或母亲)的等位基因更倾向于一起传递(分离)给子代, 这种现象称为连锁(linkage)。同一染色体上两个基因座距离越近, 则发生交换的概率越小, 于是, 两个相互连锁的基因座的重组率要小于  $\frac{1}{2}$ 。

接下来, 我们更详细地考虑两个基因座的情况。假设第一个基因座有两个等位基因 A 和 a, 第二个基因座有两个等位基因 B 和 b, 各自的等位基因频率为:

$$P(A) = p, P(a) = q, P(B) = u, P(b) = v$$

这里,  $p + q = 1, u + v = 1$ , 此时有 9 种两基因座的联合基因型。在随机婚配和无连锁的条件下, 如果父代的基因型的概率分布为:

$$\begin{array}{cccccccc} AAbb & AABb & AAAb & AaBB & AaBb & Aabb & aaBB & aaBb \\ p^2 u^2 & 2p^2 uv & p^2 v^2 & 2pqu^2 & 4pquv & 2pqv^2 & q^2 u^2 & 2q^2 uv \end{array}$$

$q^2 v^2$

那么, 子代的基因型概率与父代的基因型概率相同。因此, 如果群体的每一个基因座都处于平衡态, 且两个基因座不连锁, 则这两个基因座处于联合平衡态。对一个不处于联合平衡态的群体, 联合平衡不能在仅仅一代的随机婚配之后达到, 只有当代数  $n \rightarrow \infty$  时, 才会达到联合平衡态, 达到联合平衡的速度依赖于重组率  $\theta$ (见下面方程(3))。

如果两个基因座上的等位基因是随机关联的(独立的), 就说这两个基因座处于连锁平衡状态(linkage equilibrium)。为了说明这个概念, 我们考虑两个基因座, 各有两个等位基因: 基因座 1 上等位基因为 A 和 a, 基因座 2 上为 B 和 b, 如果这两个基因座处于连锁平衡状态, 则单体型频率满足

$$\begin{aligned} P(AB) &= P(A)P(B), P(Ab) = P(A)P(b) \\ P(aB) &= P(a)P(B), P(ab) = P(a)P(b) \end{aligned} \quad (2)$$

即在连锁平衡条件下, 两个基因座的单体型的联合概率等于相应的单个基因座上等位基因频率的乘积。如果两个基因座上的等位基因不是随机关联的, 即不独立, 这种情况就叫做等位基因关联(allelic association)或者连锁不平衡(linkage disequilibrium)。对两个

有二个等位基因的基因座,在连锁不平衡状况下,我们有:

$$\delta = P(AB) - P(A)P(B) \neq 0$$

这里  $\delta$  是偏离平衡状态的一个度量,它是连锁不平衡参数,可以证明下列关系成立:

$$P(AB) = P(A)P(B) + \delta, \quad P(Ab) = P(A)P(b) - \delta$$

$$P(aB) = P(a)P(B) - \delta, \quad P(ab) = P(a)P(b) + \delta$$

如果一个群体在初始状态下连锁不平衡( $\delta \neq 0$ ),在随机婚配条件下,当代数  $n \rightarrow \infty$  时,连锁不平衡参数将接近 0。特别地令  $\delta_0$  是起始的不平衡参数,  $\theta$  是两个基因座之间的重组率,在  $n$  代以后,有:

$$\delta_n = (1 - \theta)^n \delta_0 \quad (3)$$

这里  $\delta_n$  是  $n$  代的连锁不平衡参数。当连锁很弱,即  $\theta$  很大(接近 1/2)时,连锁不平衡参数将随着代数的增加而迅速减小。如果两个基因座不连锁,即  $\theta = 1/2$ ,则平衡状况将很快达到;如果两个基因座紧密连锁,  $\theta \approx 0$ ,则不平衡状态将持续很多代,这些是利用不平衡状况绘制基因图谱的基础。过大的连锁不平衡参数通常被视为连锁( $\theta$  很小)的证据。

想更多地了解连锁和群体遗传知识的读者可以参考 Li(1988)和 Harf、Clark(1997)写的书。大多数关于遗传统计方法的书都有基本遗传概念和术语的介绍性章节(Falconer 1988, Lange 1997; Ott 1991, Sham 1998)。Watson 等(1987)的书对分子遗传学作了一个全面的描述,Olson 等(1999)写了一篇关于复杂性状基因图谱的综述性文章。关于基本的遗传机理,读者可参考 Khoury 等(1993)和 Thompson(1986)的书。

## 第二节 分 离 分 析

孟德尔提出两等位基因在形成配子(精子或卵子)的过程中相互分离的原理,这些配子构成了下一代基因的组成物质。孟德尔分离率(segregation ratio)是指在给定婚配类型条件下,下一代的基因类型的条件概率,例如,

$$P\{aa | Aa \times Aa\} = \frac{1}{4}$$

分离率在孟德尔遗传模型下是固定不变的。因此,分离分析(segregation analysis)的任务之一就是检验观察到的后代的表现型数据是否与孟德尔遗传一致。一般来说,分离分析检验家庭数据的遗传模式。

### 一、估计等位基因频率

假定随机抽取了一个样本,并对每个个体观察了某个给定常染色体位点的表现型,如果所有的等位基因是共显性的,我们可以通过数样本中某等位基因的数目,然后除以样本中等位基因的总数来估计此等位基因的频率。每个人在给定基因座上有两个等位基因,故等位基因的总数是个体数目的两倍,例如考虑一个给定的具有两个共显性等位基因 A 和 a 的基因座,假设随机抽取  $n$  个个体来研究此基因座,这个基因座的 3 种相应的基因类型 AA、Aa、aa 的个体数目分别为  $n_{AA}$ 、 $n_{Aa}$ 、 $n_{aa}$  ( $n_{Aa} + n_{Aa} + n_{aa} = n$ ),则等位基因 A 的频率  $P_A$  可由

$$\hat{P}_A = \frac{2n_{AA} + n_{Aa}}{2n}$$

来估计,类似地,用

$$\hat{p}_a = \frac{2n_{aa} + n_{Aa}}{2n}$$

来估计 a 的频率  $P_a$ ,其中  $\hat{P}_A + \hat{P}_a = 1$ ,估计的等位基因频率的方差为:

$$Var(\hat{p}_A) = \frac{2np_A(1-p_A)}{(2n)^2} = \frac{p_A(1-p_A)}{2n}, Var(\hat{p}_a) = \frac{p_a(1-p_a)}{2n}.$$

例如,人类 MN 血型基因座上有两个等位基因 M 和 N, Li(1988)引用了 L. Ride (1935; cf. Haldane, 1938) 关于 MN 血型的数据结果:1000 多个的中国香港地区的居民接受了检查,获得了下面的数据:

血型	MM	MN	NN	总数
数目	342	500	187	1029
$\hat{P}_M = \frac{2 \times 342 + 500}{2 \times 1029} = \frac{1184}{2058} = 0.5753$				

假设在随机抽取的  $n$  个个体中,某一给定基因座上有  $k$  个共显性的等位基因,第  $i$  种等位基因的数目为  $n_i$ ,其中  $n_1 + n_2 + \dots + n_k = 2n$ ,那么,第  $i$  种等位基因频率  $p_i$  可由

$$\hat{P}_i = \frac{n_i}{2n}$$

来估计。容易看出等位基因数目  $(n_1, n_2, \dots, n_k)$  服从参数为  $(p_1, p_2, \dots, p_k)$  的多项分布,于是

$$E(\hat{p}_i) = p_i, Var(\hat{p}_i) = \frac{p_i(1-p_i)}{2n}, Cov(\hat{p}_i, \hat{p}_j) = -\frac{p_i p_j}{2n}, i \neq j$$

事实上,  $\hat{P}_i$  是  $p_i$  的极大似然估计。极大似然估计和似然比检验在遗传分析中具有重要作用。

考虑具有等位基因 A 和 a 的基因座,如果等位基因 A 是显性的,则基因型 AA 和 Aa 有相同的表现型,令  $n_A$  表示具有基因型 AA 或 Aa 的个体数目,  $n_a$  为具有基因型 aa 的个体数目,这里  $n_A + n_a = n$ 。在随机婚配的假定下,由 Hardy-Weinberg 平衡性,我们知道隐性基因型的概率等于等位基因 a 的概率的平方,即  $P(aa) = p_a^2$ ,于是有

$$\hat{p}_a^2 = \frac{n_a}{n}, \quad \hat{p}_a = \sqrt{\frac{n_a}{n}}$$

注意  $n_a$  服从参数为  $\hat{P}_a^2$  的二项分布,故  $\hat{P}_a^2$  为  $P_a^2$  的 MLE(极大似然估计)。由 MLE 的不变性原理,则  $\hat{P}_a$  为  $P_a$  的 MLE,通过  $\delta$ -方法(Rao, 1973),有

$$Var(\hat{p}_a^2) = \frac{p_a^2(1-p_a^2)}{n}, Var(\hat{p}_a) = \frac{1-p_a^2}{4n}$$

对于具有多于两个等位基因的显性基因座,例如 ABO 血型基因座,估计等位基因频率稍微复杂一点。EM 算法可用来得到等位基因频率的 MLE(Dempster 等, 1977; Lange, 1997; Little 和 Rubin, 1987; Ott, 1977; Smith, 1957)。

## 二、检验 Hardy-Weinberg 平衡

在遗传分析中,常常假定 Hardy-Weinberg 平衡。这个假定的合理性可通过 Pear-

son  $\chi^2$  统计量

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

来检验。考虑有两个等位基因的一个共显性基因座，在 Hardy-Weinberg 平衡定理假设下 ( $H_0$ )， $n$  个随机抽取的个体中，各基因型的期望数目为：

AA	Aa	aa
$nP_A^2$	$2nP_AP_a$	$nP_a^2$

令  $n_{AA}$ ,  $n_{Aa}$  和  $n_{aa}$  分别是观察到的具有基因型 AA, Aa 和 aa 的个体的数目，则

$$\hat{P}_A = \frac{2n_{AA} + n_{Aa}}{2n}, \hat{P}_a = \frac{2n_{aa} + n_{Aa}}{2n}.$$

不同基因型的期望数目可由

$$\hat{E}_{AA} = n \hat{P}_A^2, \hat{E}_{Aa} = 2n \hat{P}_A \hat{P}_a, \hat{E}_{aa} = n \hat{P}_a^2.$$

来估计，于是可得到自由度为 1 的 Pearson  $\chi^2$  检验统计量

$$\chi^2 = \frac{(n_{AA} - \hat{E}_{AA})^2}{\hat{E}_{AA}} + \frac{(n_{Aa} - \hat{E}_{Aa})^2}{\hat{E}_{Aa}} + \frac{(n_{aa} - \hat{E}_{aa})^2}{\hat{E}_{aa}}.$$

当  $\chi^2$  值大于  $x_{1-\alpha}^2(1)$  时，拒绝零假设  $H_0$ ，即人群不处于 Hardy-Weinberg 平衡状态，这里  $\alpha$  是检验的水平。

例如，下列数据来自高血压遗传研究，随机抽取 197 个个体，在血管紧张素转化酶 (ACE) 基因座上的三种基因类型的个体数目分别为：

AA	Aa	aa
26	93	78

从这个数据集，我们有：

$$\hat{P}_A = 0.3680, \hat{P}_a = 0.6320, \hat{E}_{AA} = 197 \times (0.3680)^2 = 26.68,$$

$$\hat{E}_{Aa} = 2 \times 197 \times 0.3680 \times 0.6320 = 91.63, \hat{E}_{aa} = 197 \times (0.6320)^2 = 78.69,$$

和

$$\chi^2 = \frac{(26 - 26.68)^2}{26.68} + \frac{(93 - 91.63)^2}{91.63} + \frac{(78 - 78.69)^2}{78.69} = 0.0439.$$

若检验的水平为  $\alpha = 0.05$ ，则  $\chi^2_{0.95}(1) = 3.841$ ，故检验不能拒绝群体处于 Hardy-Weinberg 平衡态的零假设。

### 三、显性基因座的分离分析

判断某个疾病是否属于单基因遗传的一个方法是验证其孟德尔分离比。考虑一个具有等位基因 A 和 a 的罕见的显性疾病基因座，假定致病的等位基因是 A，它的频率为  $P(A) = p \approx 0$ ，具有基因型 AA 和 Aa 的个体会患病，而基因型为 aa 的个体不会患病，观察到的表现型为患病或不患病。表 1 给出了随机婚配条件下，6 种婚配类型单基因显性遗传的孟德尔分离比。

在五种可能产生患病子代的婚配中，根据上面婚配类型的概率和  $p = p(A) \approx 0$  这一事实，那么最有可能出现的婚配类型是 Aa × aa。

信息量最大的抽样计划是选择有一个子代患病另一个子代不患病的家庭，这种家庭

的婚配类型通常假定为  $Aa \times aa$ 。令

表 1 随机婚配条件下, 6 种婚配类型单基因显性遗传的孟德尔分离比

婚配类型	$P(\text{婚配率})$	基因型			表现型	
		AA	Aa	aa	有病	没病
$AA \times AA$	$p^4$	1	0	0	1	0
$AA \times Aa$	$4p^3q$	$1/2$	$1/2$	0	1	0
$AA \times aa$	$2p^2q^2$	0	1	0	1	0
$Aa \times Aa$	$4p^2q^2$	$1/4$	$1/2$	$1/4$	$3/4$	$1/4$
$Aa \times aa$	$4pq^3$	0	$1/2$	$1/2$	$1/2$	$1/2$
$aa \times aa$	$q^4$	0	0	1	0	1

$\tau = P\{Aa | Aa \times aa\} = P\{\text{患病} | Aa \times aa\}$  是孟德尔分离参数, 令  $X$  表示家庭中患病的子女数目, 那么,  $X$  服从二项分布, 其概率函数为:

$$L(r | \tau) = P\{X = r\} = \binom{n}{r} \tau^r (1 - \tau)^{n-r}, \quad (4)$$

这里  $n$  是子代的数目。值得注意的是给定父母婚配类型的条件下子代的基因类型是条件独立的。下面让我们来检验孟德尔分离率, 即检验  $H_0: \tau = 1/2$ 。

1. 近似  $\chi^2$  检验 假设抽取了父母婚配类型为  $Aa \times aa$  的  $k$  个家庭, 第  $i$  个家庭有  $n_i$  个子女, 其中有  $r_i$  个子女患病 ( $n_1 + n_2 + \dots + n_k = n$ ), 令  $X = X_1 + X_2 + \dots + X_k$ , 那么  $E(X_i) = n_i \tau$ ,  $Var(X_i) = n_i \tau(1 - \tau)$ ,  $E(X) = n\tau$ ,  $Var(X) = n\tau(1 - \tau)$

由中心极限定理, 我们有

$$\frac{X - n\tau}{\sqrt{n\tau(1 - \tau)}} \xrightarrow{D} N(0, 1)$$

因此,

$$\frac{(X - n\tau)^2}{n\tau(1 - \tau)} \xrightarrow{D} \chi^2_1,$$

这里  $\xrightarrow{D}$  表示依分布收敛, 在零假设  $H_0: \tau = 1/2$  下, 我们计算检验统计量

$$\chi^2 = \frac{\left( \sum_{i=1}^k r_i - \frac{n}{2} \right)^2}{\frac{n}{4}},$$

当  $\chi^2$  较大时, 拒绝  $H_0$ 。

2. 似然比检验 由上面的家系数据结构和公式(4),  $\tau$  的似然函数为:

$$L(\tau | r_1, \dots, r_k) = \prod_{i=1}^k L(\tau | r_i) = \left[ \prod_{i=1}^k \binom{n_i}{r_i} \right] \tau^r (1 - \tau)^{n-r},$$

这里  $r = \sum_{i=1}^k r_i$ ,  $\tau$  的 MLE 为  $\hat{\tau} = r/n$ , 似然比统计量为:

$$\chi^2 = \frac{L(\hat{\tau})}{L(\frac{1}{2})} \xrightarrow{D} \chi^2_1$$

#### 四、隐性基因座的分离分析

考虑一个具有两个等位基因的罕见隐性疾病基因座。患病个体的基因型为 aa, 有三种婚配类型 Aa × Aa, Aa × aa 和 aa × aa 可能会产生患病的后代, 令  $P(A) = p$  和  $P(a) = q = 1 - p$ , 那么, 在给定一后代患病的条件下父母婚配类型的条件概率如下:

$$P(Aa \times Aa | \text{患病}) = p^2, P(Aa \times aa | \text{患病}) = 2pq, P(aa \times aa | \text{患病}) = q^2.$$

由于此疾病是罕见的隐性疾病, 即  $P(a) = q \approx 0$ , Aa × Aa 是最有可能产生患病子代的婚配类型。因此, 针对罕见的隐性遗传疾病的取样方案为选择至少一个子代患病的家庭, 并假定父母婚配类型为 Aa × Aa。对隐性遗传疾病, 婚配型 Aa × Aa 的孟德尔分离比为  $\tau = P\{aa | Aa \times Aa\} = 1/4$ 。我们可以估计和检验这个分离比。

在我们研究的样本中, 可能有某些家庭没有包括进来(虽然他们的后代也可能患病), 这时称不完全样本。Fisher(1934)在他的经典论文中认识到, 在分离分析中对不完全样本进行校正的必要性, 且提出在分析中引入确定机制。当选取至少有一个后代患病的家庭时, 最先被证实有病的后代叫做先证者(proband), 一个家庭可能有多于一个的先证者, 患病个体是先证者的概率叫做确定概率, 可表示为:

$$\pi = P\{\text{先证者} | \text{患病}\}.$$

一个有  $r$  个患病后代的家庭没有被确定的概率为:

$$P(\text{没有确定} | r \text{ 个患病}) = (1 - \pi)^r,$$

因此, 一个有  $r$  个患病后代的家庭被确定的概率是:

$$P(\text{被确定} | r \text{ 个患病}) = 1 - (1 - \pi)^r.$$

如果确定概率是 1 即  $\pi = 1$ , 那么, 后代患病的所有家庭都会被确定, 这种情况叫做完全确定(complete ascertainment);  $\pi < 1$  的情况叫做不完全确定(incomplete ascertainment); 当确定概率很小即  $\pi \approx 0$  时,  $1 - (1 - \pi)^r \approx r\pi$ , 即患病个体的家庭被确定的概率与患病子代的数目成正比。这时, 几乎所有的被确定家庭只有一个唯一的先证者, 我们称之为单一确定(single ascertainment)。

**1. 完全确定的分离分析** 在完全确定中, 每一个被确定家庭患病的子代的数目服从截尾二项分布(Fisher, 1934), 其似然函数为:

$$L(\tau | r_i, s_i) = \frac{\binom{s_i}{r_i} \tau^{r_i} (1 - \tau)^{s_i - r_i}}{1 - (1 - \tau)^{s_i}}, r_i = 1, \dots, s_i, i = 1, \dots, n \quad (5)$$

这里  $\tau = P\{aa | Aa \times Aa\}$  是孟德尔分离比,  $s_i$  是第  $i$  个家庭中子代的数目,  $r_i$  是患病的子代的数目,  $n$  是家庭的数目。分离分析的目的是估计与检验孟德尔分离比  $\tau$ 。

假定所有被确定的家庭有相同数目的子代, 即对任何  $i, s_i = s$ 。令  $a_r$  表示有  $r$  ( $r = 1, 2, \dots, s$ ) 个患病子代的家庭的数目,  $n_s$  是被确定家庭的总数, 那么,  $\sum_{r=1}^s a_r = n_s$ , 以及  $\sum_{r=1}^s r a_r = A$  是患病子代的总数。由(5), 基于  $n_s$  个被确定家庭的似然函数为:

$$L(\tau) = \prod_{i=1}^{n_s} L(\tau | r_i, s_i) = \prod_{i=1}^s \left[ \frac{\binom{s}{r} \tau^r (1 - \tau)^{s-r}}{1 - (1 - \tau)^s} \right]^{a_r}$$

$\tau$  的极大似然估计是方程  $\frac{\partial L(\tau)}{\partial \tau} = 0$  的解, 该方程等价于:

$$\frac{s\tau}{1 - (1 - \tau)^s} = \frac{A}{n_s} = \bar{r} \quad (6)$$

方程(6)没有显式解, 但我们可以用一些算法, 如通过 Newton-Raphson 方法, Fisher 计分法和 EM 算法来获得方程的近似解。 $\tau$  的 Fisher 信息量为:

$$I(\tau) = E(-\frac{\partial^2 L(\tau)}{\partial \tau^2}) = \frac{s n_s}{1 - (1 - \tau)^s} \cdot \frac{1 - (1 - \tau)^s - s\tau(1 - \tau)^{s-1}}{\tau(1 - \tau)[1 - (1 - \tau)^s]},$$

于是, MLE  $\hat{\tau}$  的方差可由  $\frac{1}{I(\hat{\tau})}$  来估计。

2. 不完全确定的分离分析 从父母婚配类型为  $Aa \times Aa$  的家庭中随机地选一个个体, 他为先证者等价于他有病且被确定, 因此,  $P\{\text{先证者}\} = P(\text{有病且被确定}) = P(\text{被确定} | \text{有病}) P(\text{有病}) = \pi\tau$ , 这里  $\pi$  指确定概率,  $\tau$  指孟德尔分离比。如果一个家庭至少有一个患病的后代, 这个家庭被认为对分离分析是有信息的, 有  $s$  个后代的家庭是有信息的且没有被确定的概率为  $(1 - \pi\tau)^s$ , 于是, 有  $s$  个后代的家庭被确定的概率为  $1 - (1 - \pi\tau)^s$ 。令  $B$  表示一个家庭中先证者数目, 那么, 一个家庭被确定等价于  $B > 0$ , 该事件的概率为  $P(B > 0) = 1 - (1 - \pi\tau)^s$ 。因此,  $r$  个后代有病的家庭被确定的似然函数为 (Morton, 1959):

$$L(\pi, \tau) = P\{X = r | B > 0; s, \pi, \tau\} = \frac{[1 - (1 - \pi\tau)^r] \binom{s}{r} \tau^r (1 - \tau)^{s-r}}{1 - (1 - \pi\tau)^s} \quad (7)$$

这个似然函数可用来估计和检验确定概率  $\pi$  和孟德尔分离率  $\tau$ 。由于这个分析涉及到谱系过程, 常称为确定偏性的校正分析。

下面给出似然函数(7)的两个特例。当确定概率  $\pi = 1$ , 即完全确定时, 似然函数(7)简化为(5)。如果确定概率非常小, 即为单一确定 ( $\pi \approx 0$ ) 时, 那么,

$$(1 - \pi)^r \approx 1 - r\pi, (1 - \pi\tau)^s \approx 1 - s\pi\tau$$

似然函数(7)可近似为:

$$L(\pi, \tau) = P\{X = r | B > 0; s, \pi, \tau\} = \binom{s-1}{r-1} \tau^{r-1} (1 - \tau)^{s-r}$$

在此例中如果被确定家庭的先证者数目知道, 似然函数由下式给出:

$$L(\pi, \tau) = P\{X = r, B = b | B > 0; s, \pi, \tau\} = \frac{\binom{r}{b} \pi^b (1 - \pi)^{r-b} \binom{s}{r} \tau^r (1 - \tau)^{s-r}}{1 - (1 - \pi\tau)^s}$$

在不完全确定条件下, 对于隐性基因座来说, 除了基于似然函数推断确定概率和孟德尔分离率之外, 还有两种简单的方法, 即先证法和单一先证者法, 现叙述如下:

先证法由 Weinberg 首先提出, Fisher(1934)作了详细的描述, 假设有  $n$  个分离的家庭被确定,  $s_i, r_i$  和  $b_i$  分别表示第  $i$  个家庭兄弟姐妹的数目、受累的后代数目和先证者的数目。那么, 先证法就是用

$$\hat{\tau} = \frac{\sum_{i=1}^n b_i(r_i - 1)}{\sum_{i=1}^n b_i(s_i - 1)} \text{ 和 } \hat{\pi} = \frac{\sum_{i=1}^n b_i(b_i - 1)}{\sum_{i=1}^n b_i(r_i - 1)}$$

来分别估计分离率和确定概率。如果被确定家庭中先证者是唯一的,这样的一个先证者叫做单一(single)先证者,令  $d$  为被确定家庭样本中单一先证者的数目,那么单一先证法就是用:

$$\hat{\tau} = \frac{\sum_{i=1}^n r_i - d}{\sum_{i=1}^n s_i - d} \text{ 和 } \hat{\pi} = \frac{\sum_{i=1}^n b_i - d}{\sum_{i=1}^n r_i - d}$$

来分别估计分离率和确定概率。

上面描述的分离分析方法是针对定性性状的(即有病或无病)。Morton 和 Maclean (1974)提出单个基因座上定量性状分离分析的“混合模型”,这种方法基于似然函数。有几个程序可以实施混合分离分析,如家系分析软件包(PAP)(Hasstedt 和 Cartwright 1987),SEGPATH (Province 和 Rao. 1995)和 POINTER(Lolouel 和 Yee 1980)。Bonney (1984)提出了一族回归的方法可以进行定量性状的分离分析,这种模型可以在调整协变量作用的同时对孟德尔模型参数进行估计,在遗传流行病统计分析(S.A.G.E)软件包中可以实施这些模型 Iston, 1986)。Terwilliger 和 Ott(1994)也提供了一系列遗传分析的计算机程序。

### 第三节 连锁分析

连锁分析考察同一染色体上两个基因座的物理距离是否相临近。两个连锁的(物理上临近的)基因座上等位基因更易于一起分离,即它们一起作为一个单位由父母传递给后代,这种现象偏离了自由组合的孟德尔第二定律。连锁分析是用来确定人类基因组上疾病易感基因位置的一种方法。人们认为,已知的标记系统和待推定的疾病基因座之间的连锁证据是此疾病由一种遗传机制造成的最有力的统计证据。连锁分析仅涉及到基因座的位置,用位置来定位基因,而不考虑此基因的生化功能。这种方法称为“定位克隆”(positional cloning)。

一个家庭中父亲(母亲)的两个基因座上等位基因由于连锁而共同分离的情况可能与另一个家庭中发生的分离情况不同。由连锁而发生的共分离现象只能在家庭内部才可以观察到,因此,考察连锁必须有家庭数据。由等位基因关联性(或连锁不平衡性)导致的共分离现象可以由一般的群体数据观察到,因此,等位基因关联性是等位基因的性质,而连锁是基因座的性质,这是两个不同但有联系的概念。

两个基因座之间的连锁用重组率  $\theta$  来衡量,两个基因座越近,发生交换的可能性越小,重组率也越小。两个极端的例子是:

(1)  $\theta = 1/2$ , 两个基因座相隔很远, 它们独立分离(即服从孟德尔自由组合的第二定律);

(2)  $\theta = 0$ , 两个基因座重合,事实上是同一个基因座。

重组率的变化范围为  $0 \leq \theta \leq 1/2$ , 通过图函数(Haldane 1919, Morgan 1928, Kosambi 1944, Rao. et al 1979), 两个基因座之间的重组率可以转换成两个基因座之间的遗传距离, 遗传距离与物理距离密切有关, 但并不相同。连锁分析估计并检验重组率  $\theta$ , 下面是

两种连锁分析的统计方法：一种基于模型，另一种与模型无关。

## 一、对数优势法

Maldane 和 Smith(1947)将极大似然比检验应用于连锁分析。在 Morton (1955) 发表了对数计分表以后，对数优势(log-odds)法得到广泛应用，该表可用来分析家庭数据。对数优势法是一种基于模型的方法，通常在对数优势法中，假定已知遗传的方式、等位基因的数目和各种基因类型的外显率， $\theta$  为唯一未知的参数。外显率(penetrance)是指给定基因类型的条件下，相应的表现型的条件概率。Ott(1999)详细地描绘了对数优势法。

给定家庭的似然函数为  $L(\theta)$ ， $\hat{\theta}$  是  $\theta$  的极大似然估计，检验  $H_0: \theta = 1/2$  及其对立假设  $H_1: \theta < 1/2$  的对数计分为：

$$Z(\hat{\theta}) = \log_{10} \frac{L(\hat{\theta})}{L(\theta = \frac{1}{2})}$$

传统上，如果  $Z(\hat{\theta}) > 3$  (Morton, 1955)，则拒绝  $H_0$ ，即声称连锁。 $Z(\hat{\theta}) > 3$  对应于一个小于  $10^{-4}$  的 P 值。

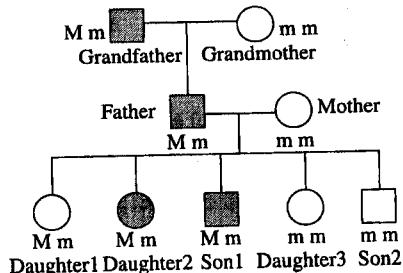


图 15-2 家系示意图

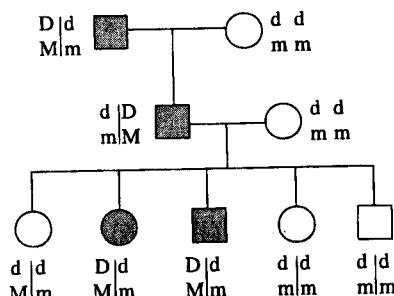


图 15-3 图 2 的连锁相

例如，图 15-2 描述了一个包含三代、有两种表现型(患病和没有患病)的家庭，且给出了每个个体的标记基因型的数据。黑色的符号表示个体患有稀有的常染色体显性遗传病，我们进一步假定这个疾病基因座有两个等位基因 D 和 d，且外显率为：

$$P(\text{有病} | DD) = P(\text{有病} | Dd) = 1, P(\text{有病} | dd) = 0.$$

从图 15-2 中的表现型和标记基因座的基因型数据，可推断出标记基因座与疾病基因座的联合基因型(即两个基因座的基因类型)及连锁相信息，它们由图 15-3 给出。因为外祖母和母亲均未患病，故他们在疾病基因座上有纯合的基因型 dd；又因为祖父患病了，故他在疾病基因座上基因型为 DD 或 Dd；又因为这种疾病很罕见，故可假定他的基因型为 Dd，这个假定对连锁分析是没有影响的，因为祖父母的基因型只用来决定父亲的连锁相。在此例中，虽然祖父有两种可能的连锁相，但每一种都导致了父亲有同一种连锁相，因为父亲患病了，故他必须至少有一个等位基因 D，同时他也必须从他母亲处得到一个等位基因 d，因此，父亲在致病基因座上的基因型为 Dd。父亲从他母亲处得到一个单体型 dm 的事实就决定了他的连锁相，因此，他的基因型如图 15-3 所示为 dm/DM。又因为母亲为双纯合的，每一个后代都从她处得到单体型 dm，于是，从父代的信息可推断出五个孩子的两

基因座上的基因型和连锁相,它们由图 15-3 给出。母亲并未提供重组的信息,父亲提供了一个重组的配子和四个非重组的配子,于是,似然函数为:

$$L(\theta) = \theta^r (1-\theta)^{N-r} = \theta(1-\theta)^4,$$

这里  $r$  表示重组的数目,  $N$  表示配子的数目,  $\theta$  的极大似然估计为  $\hat{\theta} = r/N = 0.2$ , 对数计分为:

$$Z(\hat{\theta}) = \log_{10} \frac{L(\hat{\theta})}{L(\theta=0.5)} = \log_{10} \frac{0.2 \times 0.8^4}{0.5^5} = 0.4185$$

这并不足以支持连锁的假设。这个例子处理的是连锁相已知的情况。下面的例子针对连锁相未知的情况。

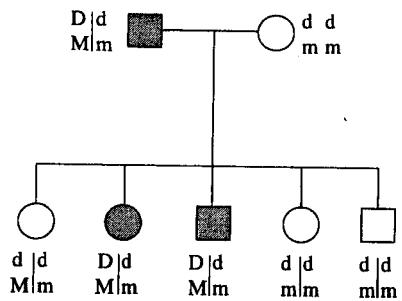


图 15-4 连锁相未知时的假定连锁相 I

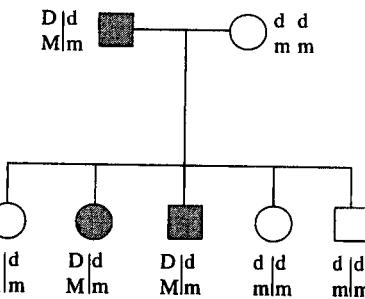


图 15-5 连锁相未知时的假定连锁相 II

例如,假设在图 15-2 和图 15-3 中缺失祖父母的信息,那么,父亲的连锁相就不能确定,父亲有基因型  $DdMm$ ,但它有两个可能的连锁相:  $dm/DM$ (连锁相 I)(图 15-4)和  $dM/Dm$ (连锁相 II)(图 15-5)。若连锁相为  $dm/DM$ ,则有一个重组体和四个非重组体的后代;若连锁相为  $dM/Dm$ ,则有一个非重组体和四个重组体的后代,两个连锁相都以  $\frac{1}{2}$  的概率出现,故似然函数为:

$$L(\theta) = P\{\text{数据}\} = P\{\text{数据} | \text{连锁相 I}\}P\{\text{连锁相 I}\} + P\{\text{数据} | \text{连锁相 II}\}P\{\text{连锁相 II}\} = \frac{1}{2}\theta(1-\theta)^4 + \frac{1}{2}\theta^4(1-\theta)$$

在这个例子中,  $\theta$  的极大似然估计不再具有显式解。很多数值程序可用来获得  $\theta$  的极大似然估计( $\hat{\theta}$ )。若用 LINKAGE 程序(Lathrop, et al 1990),我们得到( $\hat{\theta}$ )=0.21(近似值),对数计分  $Z(\hat{\theta})$  为 0.1249。Terwiegger 和 Ott(1994)给出了 LINKAGE 程序的详细描述和使用方法。

如果有多个家庭的数据,假定各个家庭的数据是独立的,则联合似然函数为各个家庭的似然函数的乘积。对数优势法可以类似地使用。

## 二、受累同胞对(ASP)方法

同胞对是最简单的家庭单位且容易确定,故同胞对连锁分析被广泛用于复杂性状遗传机制的研究。Penrose(1935)首次提出这种方法,他当时是基于这样一种思想:有相同表现型的同胞对拥有共同等位基因的概率较大,而不同表现型的同胞对应该没有共同的等位基因。这种思想的进一步发展就导致了受累同胞对(ASP)法。ASP 方法确定一对患

病的同胞，他们拥有共同易感等位基因的概率较大，共同易感等位基因的一个衡量标准是血源一致( IBD ) 等位基因的数目。由同一祖先的同一等位基因遗传下来的两个等位基因即为一个 IBD，例如，父母的婚配类型为  $Aa \times aa$ ，且同胞对的基因型均为  $Aa$ ，则此同胞对的等位基因‘A’为 IBD，但等位基因‘a’则可能是也可能不是。

令  $I$  表示同胞对的 IBD 等位基因的数目，它是一个随机变量，其分布为：

$$P(I=0) = \frac{1}{4}, P(I=1) = \frac{1}{2}, P(I=2) = \frac{1}{4},$$

因此，

$$E(I) = 1, Var(I) = \frac{1}{2}$$

给定同胞对中成员患病状况的条件下， $I$  的条件分布提供了 ASP 方法的理论基础，Suarez 等(1978)在有两个等位基因的基因座假定下，得出了这个分布，Risch(1989, 1990)把它推广到其他亲属关系和复基因座模型。在描述 ASP 方法之前，我们来介绍一些符号和概念。

假定在性状基因座上有等位基因 T 和 t，且  $P(T) = P, P\{t\} = q$ 。令  $Y$  为一取两值的随机变量，它表示个体是否感染某一病症，即

$$Y = \begin{cases} 1, \text{患病} \\ 0, \text{未患病} \end{cases}$$

假定三个外显率函数为：

$$f_1 = P\{Y=1|TT\} = P\{\text{患病}|TT\}, f_2 = P\{Y=1|Tt\}, f_3 = P\{Y=1|tt\}$$

在 Hardy-Weinberg 平衡下，群体中此疾病的流行率为：

$$\begin{aligned} K_p &= P\{Y=1\} = P\{Y=1|TT\}P\{TT\} + P\{Y=1|Tt\}P\{Tt\} + P\{Y=1|tt\}P\{tt\} \\ &= p^2 f_1 + 2pqf_2 + q^2 f_3 \end{aligned}$$

这种基因座上二分类性状的方差为：

$$V_G = Var(E(Y|G)) = V_A + V_D$$

这里  $G$  是表示三种基因类型的随机变量， $V_A = 2pq[p(f_2 - f_1) + q(f_3 - f_2)]$  是可加性方差， $V_D = p^2q^2(f_1 - 2f_2 + f_3)^2$  是显性方差。可加性模型对应于  $f_2 = \frac{f_1 + f_3}{2}$ ，显性模型对应于  $f_1 = f_2 = 1$  和  $f_3 = 0$ ，隐性模型对应于  $f_1 = f_2 = 0$  和  $f_3 = 1$ 。

令  $I_M$  表示在标记基因座上同胞对具有 IBD 等位基因的数目， $X$  表示同胞对中患病的数目， $\theta$  为性状基因座和标记基因座之间的重组率。Suarz 等(1978)得出分布函数  $P\{I_M=j|X=k\}, j=0,1,2, k=0,1,2$ ，见表 2。

这里

$$d_2 = 4(K_P^2 + V_A/2 + V_D/4),$$

$$d_1 = 4(2K_P - V_A - V_D/2 - 2K_P^2),$$

$$d_0 = 4(1 - 2K_P + K_P^2 + V_A/2 + V_D/4)$$

$$\Psi = \theta^2 + (1 - \theta)^2$$

表 2 分布函数  $P\{I_M = j | X = k\}, j = 0, 1, 2, k = 0, 1, 2,$ 

j=2	j=1	j=0
$k=2 \frac{1}{4} + \frac{(\Psi - \frac{1}{2})V_A + (\Psi^2 - \frac{1}{4})V_D}{d_2}$	$\frac{1}{2} - \frac{2(\Psi^2 - \Psi + \frac{1}{4})V_D}{d_2}$	$\frac{1}{4} - \frac{(\Psi - \frac{1}{2})V_A + (2\Psi - \Psi^2 - \frac{3}{4})V_D}{d_2}$
$k=1 \frac{1}{4} - \frac{(2\Psi - 1)V_A + (2\Psi^2 - \frac{1}{2})V_D}{d_1}$	$\frac{1}{2} + \frac{2(2\Psi^2 - 2\Psi + \frac{1}{2})V_D}{d_1}$	$\frac{1}{4} + \frac{(2\Psi - 1)V_A + (4\Psi - 2\Psi^2 - \frac{3}{2})V_D}{d_1}$
$k=0 \frac{1}{4} + \frac{(\Psi - \frac{1}{2})V_A + (\Psi^2 - \frac{1}{4})V_D}{d_0}$	$\frac{1}{2} - \frac{2(\Psi^2 - \Psi + \frac{1}{4})V_D}{d_0}$	$\frac{1}{4} - \frac{(\Psi - \frac{1}{2})V_A + (2\Psi - \Psi^2 - \frac{3}{4})V_D}{d_0}$

从表 2 可看出在  $X=2$  时, 对立假设下  $I_M$  的条件分布与零假设下  $I_M$  的条件分布( $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ )偏离的程度是三种情况( $X=2, X=1, X=0$ )中最大的, 因此,  $X=2$  的设计对考察连锁提供最多的信息。有许多基于 ASP 数据考察连锁的检验统计量, 最流行的一个为均值检验, 它的定义是:

$$T = \frac{\left(n_2 + \frac{1}{2}n_1\right) - \frac{n}{2}}{\left(\frac{n}{8}\right)^2}$$

这里  $n$  指受累同胞对的总数,  $n_2$  表示  $I_M=2$  的受累同胞对的数目,  $n_1$  表示  $I_M=1$  的受累同胞对的数目。在无连锁的零假设下,  $T$  有近似标准正态分布。给定( $K_P, V_A, V_D, \theta$ )的值, 功效和样本大小的计算可由表 2 的有关概率值得出。Knapp 等(1995), Suarz 和 Eerdewegh(1984), Blaekweleder 和 Elston(1985)研究了均值检验统计量的功效, 表明了在各种情况下功效都不错。

Risch(1990)推导了基于复发风险的 ASP 方法。同胞对的复发风险定义为同胞对中一个患病的条件下另一个患病的条件概率, 即  $K_R = P(Y_2=1 | Y_1=1)$ , 这里  $Y_1$  和  $Y_2$  表示同胞对的患病状况。同胞对的复发风险率为同胞对的复发风险与群体发病率的比, 即  $\lambda_S = \frac{K_R}{K_P}$ 。同样, 父母与子女的复发风险率可通过在上式中用父母与子女的复发风险替代而得, 记为  $\lambda_O$ 。给定同胞对中两个均患病的条件下,  $I_M$  的条件分布为:

$$\begin{aligned} z_0 &= \frac{1}{4} - \frac{1}{4\lambda_S}(2\Psi - 1)[(\lambda_S - 1) + 2(1 - \Psi)(\lambda_S - \lambda_O)] \\ z_1 &= \frac{1}{2} - \frac{1}{2\lambda_S}(2\Psi - 1)^2(\lambda_S - \lambda_O) \\ z_2 &= \frac{1}{4} + \frac{1}{4\lambda_S}(2\Psi - 1)[(\lambda_S - 1) + 2\Psi(\lambda_S - \lambda_O)] \end{aligned}$$

其中  $z_i = P(I_M = i | \text{同胞对中两个均患病})$  和  $= \theta^2 + (1 - \theta)^2$ 。注意这里的参数系统与 Suarz 等(1978)的参数系统不同, 但对于零假设  $H_0: \theta = \frac{1}{2}$  和对立假设  $H_1: \theta < \frac{1}{2}$  的均值检验统计量和前面一样, 不同的是根据参数( $\lambda_S, \lambda_O, \theta$ )计算功效和样本大小。具有参数( $\lambda_S, \lambda_O, \theta$ )的三项分布的似然比检验也能进行。受累同胞对分析方法分别被 Risch

(1990) 和 Weeks、Lange(1988) 推广到患病亲属对和患病亲属集或者说患病家族成员(APM)。

### 三、Haseman-Elston 过程

上面的对数优势法和 ASP 方法处理的是定性性状的基因连锁问题, 然而, 很多复杂性状都是定量的, 也就是说, 他们取连续的值。是否患疾病经常是由数量来衡量的。例如, 高血压由血压的高低来决定, 肥胖症是由体重来决定。因此, 研究标记基因座和定量性状基因座之间的连锁分析统计方法很重要。

令  $x_{1j}$  和  $x_{2j}$  表示同胞对中两同胞的连续性状的值。我们假定性状有如下结构:

$$x_{1j} = \mu + g_{1j} + e_{1j}$$

$$x_{2j} = \mu + g_{2j} + e_{2j}$$

这里  $\mu$  为总体均值,  $g_{1j}$  和  $g_{2j}$  表示基因对性状值的贡献,  $e_{1j}$  和  $e_{2j}$  为残差。

我们考虑只有两个等位基因  $A_1$  和  $A_2$  的单基因座情况,  $A_1$  和  $A_2$  的频率分别为  $p$  和  $q = 1 - p$ 。对三种可能的不同基因型, 下表给出了个体的平均性状值:

$A_2A_2$	$A_2A_1$	$A_1A_1$
$-a$	$d$	$a$

于是可加性遗传方差为  $\sigma_a^2 = 2pq[a + (q - p)d]^2$ , 显性方差为  $\sigma_d^2 = (2pqd)^2$ , 总的遗传方差为可加性方差和显性方差的和, 即  $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$ 。令  $\sigma_e^2 = E(e_{1j} - e_{2j})^2$  以及  $\rho$  为同胞对间残差的相关系数。在本节中, 我们假定没有显性方差, 即  $\sigma_d^2 = 0$ 。

Haseman 和 Elston(1972) 表明了在给定同胞对共享性状基因座上 IBD 等位基因率  $\pi_j$  的条件下, 同胞对性状值平方差的条件期望满足回归方程

$$E(y_j | \pi_j) = \sigma_e^2 + 2\sigma_g^2 - 2\sigma_g^2\pi_j = \beta_0 + \beta_1\pi_j, \quad (8)$$

这里  $y_j = (x_{1j} - x_{2j})^2$ ,  $\beta_0 = \sigma_e^2 + 2\sigma_g^2$  和  $\beta_1 = 2\sigma_g^2$

对于候选基因情况, 由同胞对数据  $(y_1, \pi_1), \dots, (y_n, \pi_n)$  和回归方程(8), 我们可获得  $\sigma_e^2$  和  $\sigma_g^2$  的最小二乘估计  $\hat{\sigma}_e^2$  和  $\hat{\sigma}_g^2$ , 则遗传率(heritability)的估计可通过用  $\hat{\sigma}_e^2$  和  $\hat{\sigma}_g^2$  取代公式:

$$H = \frac{\sigma_g^2}{\sigma_e^2 + \sigma_g^2}$$

中  $\sigma_e^2$  和  $\sigma_g^2$  而得到。这种方法已经推广到多个等位基因的复基因座上(Stoesz, et al, 1997)。

如果用标记基因座上 IBD 等位基因率  $\pi_{jm}$  代替性状基因座上 IBD 等位基因率  $\pi_j$ , 那么回归方程变为(Haseman 和 Elston, 1972):

$$E(y_j | \pi_{jm}) = \sigma_e^2 + 2\sigma_g^2\Psi - 2(1-2\theta)^2\sigma_g^2\pi_{jm} = \gamma_0 + \gamma_1\pi_{jm}, \quad (9)$$

这里  $y_j = (x_{1j} - x_{2j})^2$ ,  $\gamma_0 = \sigma_e^2 + 2\sigma_g^2\Psi$ ,  $\Psi = \theta^2 + (1-\theta)^2$ ,  $\gamma_1 = -2(1-2\theta)^2\sigma_g^2$ ,  $\theta$  为标记基因座和性状基因座之间的重组率, 第  $j$  个同胞对在标记基因座上 IBD 等位基因率  $\pi_{jm}$  的取值为 0, 1/2 和 1。 $\pi_{jm}$  经常可估计出来, 用估计值  $\hat{\pi}_{jm}$  表达的回归方程为:

$$E(y_j | \hat{\pi}_{jm}) = \gamma_0 + \gamma_1 \hat{\pi}_{jm}, \quad (10)$$

这里  $\gamma_0$  和  $\gamma_1$  与(9)式中的意义相同,  $\hat{\pi}_{jm}$  取值  $\frac{i}{4}$ ,  $i = 0, 1, 2, 3, 4$  (Haseman 和 Elston,

1972)。注意到当  $\sigma_g^2 > 0$  时,由  $\gamma_1 = 0$  推出  $\theta = \frac{1}{2}$ 。 $\gamma_1$  的最小二乘估计  $\hat{\gamma}_1$  可由同胞对遗传数据  $(y_1, \hat{\pi}_{1m}), \dots, (y_n, \hat{\pi}_{nm})$  和回归方程(10)而得。我们可用  $\hat{\gamma}_1$  来检验  $H_0: \theta = \frac{1}{2}$  (无连锁)与对立假设  $H_1: \theta < \frac{1}{2}$ 。因此,若  $\hat{\gamma}_1$  明显为负,则表示  $\theta < \frac{1}{2}$ ,即在水平  $\alpha$  的检验中若  $\hat{\gamma}_1$  小于适当的值  $C < 0$ ,则拒绝  $H_0$ 。

#### 四、ED 和 EC 同胞对设计

Haseman-Elston(1972)的模型基于随机抽取的同胞对,但有些基于性状值抽取同胞对的抽样方案具有更大的功效(Blacker 和 Elston(1982), Cary 和 Williamson(1991), Eaves 和 Meyer(1994), Risch 和 Zhang(1995))。Risch 和 Zhang(1995)给出了基于性状值抽取同胞对的三类有效方法,为定量性状基因座(QTL)的连锁分析提供了较大的功效:①极端相异(ED)的同胞对:即其中一个具有大的性状值,而另一个具有小的性状值;②大性状值的极端一致(EC)同胞对;③小性状值的极端一致同胞对。Risch 和 Zhang(1995), Risch 和 Zhang(1996), Zhang 和 Risch(1996), Zhao, Zhang 和 Ratler(1997)在不同遗传模型下考察三种同胞对的功效,得出的结论是极端相异的同胞对具有最大的功效,因此,在进行人类 QTL 基因座连锁分析时他们推荐使用极端相异的同胞对设计。Eaves 和 Meyer(1994)也通过类似的方法,得到 ED 同胞对的功效。

为了作基因类型和基因连锁研究,假定确定了  $N$  个 ED 同胞对,令  $n_0, n_1$  和  $n_2$  是分别具有 IBD=0,1 和 2 的同胞对的数目。若标记基因座与性状基因座是连锁的,则会有较多的 ED 同胞对具有 IBD=0,这是由抽样方案所决定的。因此,如果  $n_0$  明显大于  $n_2$ ,则表明有连锁;在无连锁的  $H_0$  下,( $n_0, n_1, n_2$ )服从参数为  $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$  的三项分布,于是可以建立一个基于  $n_0 - n_2$  的统计量,由于

$$E_{H_0}(n_0 - n_2) = 0, \text{Var}_{H_0}(n_0 - n_2) = \frac{N}{2}$$

我们定义检验统计量

$$T_{ED} = \frac{n_0 - n_2}{\sqrt{\frac{N}{2}}},$$

这个统计量近似服从标准正态分布。当  $T_{ED}$  足够大时拒绝  $H_0$ ,即可宣称两基因座是连锁的。样本大小和功效的公式已由 Risch 和 Zhang(1995)给出。对于 EC 同胞对设计,检验统计量为:

$$T_{EC} = \frac{n_2 - n_0}{\sqrt{\frac{N}{2}}}$$

现有的几个将 ED 和 EC 同胞对合二为一的检验方法给每一个 ED(EC)同胞对加相等的权重(Gu 等(1996), Li 和 Zhang(2000))。Rao(1998)注意到通过给性状分布的尾部加不同的权重,可以改进上面的方法。Li 和 Gastwirth(2000)通过给相异程度高的 ED 同胞对和大性状值的 EC 同胞对加更大权重的方法发展了这一检验。

## 五、传递不平衡检验(TDT)

假定在一个疾病基因座上有两个等位基因  $D_1$  和  $D_2$ , 在标记基因座上有两个等位基因  $M_1$  和  $M_2$ 。假设确定了  $n$  个患病的子女, 他们分别来自  $n$  个不同的家庭。在这  $n$  个家庭中, 父母将有  $4n$  个标记基因, 其中  $2n$  个传递给了下一代, 另外  $2n$  个没有传递。若标记基因座在疾病基因座的附近, 且疾病等位基因源于最近的一次基因突变, 那么, 与疾病等位基因相关联的标记等位基因将以更高的频率出现在患病的个体中(相对于正常个体而言), 这个关联的标记等位基因相对于另一个标记等位基因的不平衡传递表明了标记基因座和疾病基因座之间存在连锁。

表 3 摘自 Spielman 等(1993), 它概括了  $2n$  个父母传递给后代的等位基因的数目和没有传递的数目。

表 3  $n$  个后代的  $2n$  个父母传递和没有传递标记等位基因  $M_1$  和  $M_2$  的数目

传递的等位基因	没有传递的等位基因		总数
	$M_1$	$M_2$	
$M_1$	$a$	$b$	$a + b$
$M_2$	$c$	$d$	$c + d$
总数	$a + c$	$b + d$	$2n$

注意到在上面的  $2 \times 2$  表中  $b$  代表在标记基因座上基因型为  $M_1M_2$ , 传递给后代  $M_1$ 、而没有传递  $M_2$  的父母的数目。因为上表中每一个患病后代的父母提供了一个传递的等位基因和一个没有传递的等位基因, 故 Spielman 等 1993 年(Spielman 和 Ewen 1996)提出的传递不平衡检验(TDT)是由婚配的对照设计导出的 McNemar 检验, 其检验统计量为:

$$\chi^2_{TDT} = \frac{(b - c)^2}{b + c}$$

McNemar 检验基于二项分布对标准正态分布的近似, 于是, 该检验统计量近似服从于自由度为 1 的  $\chi^2$  分布。TDT 的理论背景由表 4 给出, 它摘自 Curnow 等(1998)。

表 4  $n$  个后代的  $2n$  个父母传递和不传递标记等位基因  $M_1$  和  $M_2$  的概率

传递了的等位基因	没有传递的等位基因	
	$M_1$	$M_2$
$M_1$	$m^2 + \frac{Bm\delta}{P}$	$m(1-m) + \frac{B(1-\theta-m)\delta}{P}$
$M_2$	$m(1-m) + \frac{B(\theta-m)\delta}{P}$	$(1-m)^2 - \frac{B(1-m)\delta}{P}$

在表 4 中,  $m = P(M_1)$  和  $P = P(D_1)$  是等位基因频率,  $\theta$  是标记基因座和性状基因座之间的重组率,  $\delta = P(M_1D_1) - P(M_1)P(D_1)$  是连锁不平衡(关联)参数以及

$$B = \frac{p[f(f_{11} - f_{12}) + (1-p)(f_{12} - f_{22})]}{p^2 f_{11} + 2p(1-p)f_{12} + (1-p)^2 f_{22}}$$

其中  $f_{11} = P(\text{患病} | D_1D_1)$ ,  $f_{12} = P(\text{患病} | D_1D_2)$ , 和  $f_{22} = P(\text{患病} | D_2D_2)$  是三个外显率,

即个体具有基因型  $D_1D_1, D_1D_2$  和  $D_2D_2$  而患病的概率。表 4 中右上角的项  $m(1-m) + \frac{B(1-\theta-m)\delta}{P}$  是在给定子女患病的条件下, 父母的基因型为  $M_1M_2$  且传递了  $M_1$  的条件概率, 即

$$P_{12} = P(\text{父母} = M_1M_2 \rightarrow M_1 | \text{子女患病}) = m(1-m) + \frac{B(1-\theta-m)\delta}{P}.$$

类似地, 表中左下角的概率为:

$$P_{21} = P(\text{父母} = M_1M_2 \rightarrow M_2 | \text{子女患病}) = m(1-m) + \frac{B(\theta-m)\delta}{P}.$$

若  $\delta \neq 0$ (关联), 则  $H_0: \theta = \frac{1}{2}$  等价于  $H_0: P_{12} = P_{21}$ , 因此, TDT 检验是一个连锁和关联的联合检验。表 4 中主对角线上的项与重组率无关(独立), 这与直观认识是一致的, 即父母为纯合子便没有连锁的信息, 这就是为什么表 3 中 TDT 统计量  $\chi^2_{\text{TDT}}$  只包括了表 3 中的  $b$  和  $c$  而不包括  $a$  和  $d$  的原因。

TDT 方法与单体型相关风险(HRR)(Falk 和 Ruhinstein(1987); Ott(1989))有关。它有一个优点, 即它只需要有子女患病的家庭的父母和后代的数据, 它不像 ASP 方法需要多个兄弟姐妹的数据。它的一个缺点是只能考查存在关联时的连锁。TDT 方法已经被推广到有多个标记基因座的情况(Bickeböller 和 Clerget-Darpoux(1991); Sham 和 Curtis(1995); Schard(1996); Spielman 和 Ewens(1996))、没有父母基因型信息的情况(Curtis(1997); Boehnke 和 Langefeld(1998); Monks 等(1998); Schaid 和 Rowland(1998); Spielman 和 Ewens(1998))和定量性状基因座的情况(Allison(1997); Rabinowitz(1997); Schaid 和 Rowland(1999))。

## 第四节 讨 论

这章所讨论的统计方法仅仅涉及到分子生物学数据所产生问题的一小部分, 这些分子生物学数据适合于寻找疾病基因。当前使用的各种统计方法的性质还需要继续研究, 新的统计方法需要不断发展。要了解在上世纪中统计学对遗传学的主要贡献以及当前和以后的研究方向, 读者可参阅 Elson 和 Thompson(2000)的文章, 它是这方面的一篇优秀的综述文章。

致谢: 这项工作得到了美国国家癌症研究所基金 CA 64363 和中国自然科学基金 19941003 的部分资助。我们感谢 Chin Long Chiang 教授和 Dennis Buckman 先生的极高评价。

## 参 考 文 献

- Allison D. B. (1997). Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* 60, 676-690.
- Bickeböller H. and Clerget-Darpoux F. (1995). Statistical properties of the allelic and genotypic transmission/disequilibrium test for multi-allelic markers. *Genet. Epidemiol.* 12, 865-870.

3. Blackwelder W. C. and Elston R. C. (1982). Power and robustness of sib-pair linkage test and extension to larger sibships. *Commun. Stat. Theory Methods* 11, 449-484.
4. Boehnke M. and Langefeld C. D. (1998). Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am. J. Hum. Genet.* 62, 950-961.
5. Bonney G. E. (1984). On the statistical determination of major gene mechanisms in continuous human traits: Regressive models. *American Journal of Medical Genetics* 18, 731-749.
6. Carey G. and Williamson J. (1991). Linkage analysis of quantitative traits: increased power by using selected samples. *Am. J. Hum. Genet.* 49, 786-796.
7. Curnow R. N., Morris A. P., and Whittaker J. C. (1998). Locating genes involved in human disease. *Appl. Statist.* 47, 63-76.
8. Curtis D. (1997). Use of siblings as controls in case-control association studies. *Ann. Hum. Genet.* 61, 319-333.
9. Dempster A. P., Laird N. M., and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc. B* 39, 1-38.
10. Eaves L. and Meyer J. (1994). Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotyping. *Behav. Genet.* 24, 443-455.
11. Elston R. C., Bailey-Watson J. E., Bonney G. E., et al (1986). A package of computer programs to perform statistical analysis for genetic epidemiology. Presented at the Seventh International Congress of Human Genetics, Berlin.
12. Elston R. C. and Thompson E. A. (2000). A century of biometrical genetics. *Biometrics* 56, 659-666.
13. Falconer D. S. (1989). Introduction to quantitative genetics. Longman Scientific & Technical with John Wiley & Sons, Inc. New York.
14. Falk C. T. and Rubinstein P. (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* 51, 227-233.
15. Fisher R. A. (1952). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics* 6, 13-25.
16. Gu C., Todorov A., and Rao D. C. (1996). Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of QTLs. *Genet. Epidemiol.* 13, 513-533.
17. Haldane J. B. S. (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *J. Genet.* 8, 299-309.
18. Haldane J. B. S. and Smith C. A. B. (1947). A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics* 14, 10-31.
19. Hardy G. H. (1908). Mendelian proportions in a mixed population. *Science* 28, 49-50.
20. Hartl D. L. and Clark A. G. (1997). Principles of population genetics. Sinauer Associates, Inc. Sunderland, Massachusetts.
21. Haseman J. K. and Elston R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* 2, 3-19.
22. Hasstedt S. J. and Carwright P. E. (1981). PAP-pedigree analysis package, University of Utah, Department of Medical Biophysics and Computing, Technical Report No. 13. Salt Lake City, Utah.
23. Khoury M. J., Beaty T. H., and Cohen B. H. (1993). Fundamentals of genetic epidemiology. Oxford University Press, New York and Oxford.
24. Knapp M., Wassmer G., and Baur M. P. (1995). Linkage analysis in nuclear families, I. Optimality criteria for affected sib-pair tests. *Human Heredity* 44, 37-43.

25. Kosambi D. D. (1994). The estimation of map distances from recombination values. *Ann. Eugen.* 12, 172-175.
26. Lalouel J. M. and Yee S. (1980). POINTER: a computer program for complex segregation analysis with pointers. Technical Report, Population Genetics Laboratory, University of Hawaii, Honolulu.
27. Lange K. (1997). Mathematical and statistical methods for genetic analysis. Springer - Verlag, New York.
28. Lathrop G. M. , Lalouel J. M. , Juier C. , et al. (1984). Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. USA* 81, 3443-3446.
29. Li C. C. (1988). First course in population genetics. The Boxwood Press, Pacific Grove, California.
30. Li Z. and Zhang H. (2000). Mapping quantitative trait loci in humans using both extreme discordant and concordant sib pairs: a unified approach for meta-analysis. *Commun Stat Theory Methods* 29, 1115-1127.
31. Li Z. and Gastwirth J. L. (2001). A weighted test using both extreme discordant and concordant sibpairs for detecting linkage. *Genetic Epidemiology*, In Press.
32. Little R. J. A. and Rubin D. B. (1987). Statistical analysis with missing data. Wiley, New York.
33. Monks S. A. , Kaplan N. L. , and Weir B. S. (1998). A comparative study of sibship tests of linkage and/or association. *Am. J. Hum. Genet.* 63, 1507-1516.
34. Morgan T. H. (1928) The theory of genes. Yale University Press, New Haven, Conn.
35. Morton N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics* 7, 277-318.
36. Morton N. E. (1959). Genetic tests under incomplete ascertainment. *American Journal of Human Genetics* 11, 1-16.
37. Morton N. E. and MacLean C. J. (1974). Analysis of family resemblance. III . Complex segregation analysis of quantitative traits. *American Journal of Human Genetics* 26, 489-503.
38. Olson J. M. , Witte J. S. , and Elston R. C. (1999). Tutorial in biostatistics: genetic mapping of complex traits. *Statistics in Medicine* 18, 2961-2981.
39. Ott J. (1977). Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. *Ann. Hum. Genet.* 40, 443-454.
40. Ott J. (1989). Statistical properties of the haplotype relative risk. *Genet. Epidemiol.* 6, 127-130.
41. Ott J. (1999). Analysis of human genetic linkage. The Johns Hopkins University Press. Baltimore and London.
42. Penrose L. S. (1935). The detection of autosomal linkage in data which consist of pairs brothers and sisters of unspecified parentage. *Ann. Eugen.* 6, 133-138.
43. Penrose L. S. (1953). The general purpose sib-pair linkage test. *Ann. Eugen.* 18, 120-124.
44. Province M. A. and Rao D. C. (1995). General purpose model and a computer program for combined segregation and path analysis (SEGPATH): automatically creating computer programs from symbolic language model specifications. *Genet. Epidemiol.* 12, 203-219.
45. Rabinowitz D. (1997). A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.* 47, 342-350.
46. Rao D. C. (1998). CAT scans, PET scans, and genomic scans. *Genetic Epidemiology* 15, 1-18.
47. Rao D. C. , Keats B. J. B. , Lalouel J. M. , et al (1979). A maximum likelihood map of chromosome 1. *Am. J. Hum. Genet.* 31, 680-96.
48. Risch N. (1990). Linkage strategies for genetically complex traits II . The power of affected relative pairs. *Am. J. Hum. Genet.* 46, 229-241.
49. Risch N. and Zhang H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans.

- Science 268, 1584-1589
50. Risch N. and Zhang H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: sample size considerations. Am. J. Hum. Genet. 58, 836-843
51. Schaid D. J. (1996). General score tests for associations of genetic markers with disease using cases and parents. Genet. Epidemiol. 13, 423-449.
52. Schaid D. J. and Rowland C. (1998). The use of parents, sibs, and unrelated controls to detection of association between genetic markers and diseases. Am. J. Hum. Genet. 63, 1492-1506.
53. Schaid D. J. and Rowland C. M. (1999). Quantitative trait transmission disequilibrium test: allowance for missing parents. Genet. Epidemiol. 17, S307- S312.
54. Sham P. (1998). Statistics in human genetics. Arnold, London and New York.
55. Sham P. C. and Curtis D. (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. Ann. Hum. Genet. 59, 323-336.
56. Smith C. A. B. (1957). Counting methods in genetical statistics. Ann. Hum. Genet. 21, 254-276.
57. Spielman R. S. and Ewens W. J. (1996). The TDT and other family-based tests for linkage disequilibrium and association. Am. J. Hum. Genet. 59, 983-989.
58. Spielman R. S. and Ewens W. J. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am. J. Hum. Genet. 62, 450-458.
59. Spielman R. S., McGinnis R. E., and Ewens W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). American Journal of Human Genetics 52, 506-516.
60. Stoesz M. R. , Cohen J. C. , Mooser V. , et al (1997). Extension of the Haseman-Elston method to multiple alleles and multiple loci: theory and practice for candidate genes. Ann. Hum. Genet. 61, 263-274.
61. Suarez B. K. , Rice J. , Reich T. (1978). The generalized sib pair IBD distribution: its use in the detection of linkage. Ann. Hum. Genet. 42, 87-94.
62. Suarez B. K. and Van Eerdewegh P. (1984). A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. American Journal of Medical Genetics 18, 135-146.
63. Terwilliger J. D. and Ott J. (1994). Handbook of human genetic linkage. The John Hopkins University Press, Baltimore and London.
64. Thompson E. A. (1986). Genetic epidemiology: a review of the statistical basis. Statistics in Medicine 5, 291-302.
65. Watson J. D. , Hopkins N. H. , Roberts J. W. , et al (1987). Molecular biology of the gene. The Benjamin/Cummings Publishing Company, Inc. , Menlo Park, California.
66. Weeks D. E. and Lange K. (1988). The affected-pedigree-member method of linkage analysis. Am. J. Hum. Genet. 42, 315-326.
67. Weinberg W. (1908). Über den Nachweis der Vererbung beim Menschen. Jahresh. Verein f. vaterl. Naturk. in Wurttemberg 64, 368-82.
68. Wentworth E. N. and Remick B. L. (1916). Some breeding properties of the generalized Mendelian population. Genetics 1, 608-616.
69. Zhang H. and Risch N. (1996). Mapping quantitative trait loci in humans using extreme concordant sib pair: selected sampling by parental phenotypes. Am. J. Hum. Genet. 59, 951-957
70. Zhao H. , Zhang H. , and Rotter J. I. (1997). Cost-effective sib-pair designs in the mapping of quantitative-trait loci. Am. J. Hum. Genet. 60, 1211 -1221.

**第一作者简介:**李照海,理学博士 现为美国乔治华盛顿大学统计系生物统计中心统计与生物统计副教授。于华中师范大学获得数理统计硕士学位,在美国哥伦比亚大学获得统计学博士学位,他的研究涉及遗传流行病学中的统计方法及 Meta 分析的经典贝叶斯方法。